UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

Transformação de geometria de redes complexas para espaços lineares usando observadores comóveis

Rui Filipe Gomes Figueiredo

Mestrado Integrado em Engenharia Física

Dissertação orientada por:

Professor Doutor João Carlos Caetano de Freitas Pires da Cruz

Professor Doutor Nuno Miguel Azevedo Machado de Araújo

2021

Agradecimentos

Quero agradecer aos meus orientadores Professor João Pires da Cruz e Professor Nuno Araújo por terem aceite ajudar-me neste passo fundamental da minha vida e pela paciência enorme que demonstraram durante a escrita desde documento. Agradeço também ao Tiago Barroca pela instrução em Python e ao Bruno Pereira a extensa revisão do texto final. Finalmente, um especial obrigado à Closer Consulting pelos desafios que me propõe todos os dias.

Canta per me addio
quel dolce suono

Yuki Kajiura

Resumo

Dois conceitos fundamentais em Finanças são o **ativo** e o **passivo** de uma entidade. O que tenho (ativo) e o que devo (passivo) são conceitos que fazem parte do senso comum e acredita-se que são indicadores da "saúde económica" de empresas e indivíduos. Assim, muitas foram as tentativas de extrair informações e previsões destes valores económicos. No entanto, "medir a economia" é um desafio que não deve ser menosprezado. Se é fácil saber a quantidade de moeda em circulação, esta é apenas uma representação da economia; o que nos interessa medir verdadeiramente são as trocas de bens e serviços e o "trabalho" que as tornou possíveis (aqui, trabalho refere-se ao conceito humano de trabalhar em vez do usado na física). –Junta-se a isto a natureza altamente confidencial e pessoal de todos os dados económicos.

Um ponto que tem sido especialmente difícil de ultrapassar, e que muitas vezes é desprezado em análises económicas, é que a distribuição de uma variável económica (riqueza, dívida, rendimentos, etc) não é Gaussiana. De facto, a variância destas distribuições é muitas vezes infinita, ganhando por isso a designação de distribuições *heavy-tail* na literatura económica. O facto da variância ser infinita invalida a aplicação do Teorema do Limite Central (TLC), o que significa que o valor dos momentos não converge com o aumento da quantidade de dados.

O presente documento estabelece um algoritmo de transformação que, aplicado a uma distribuição com variância infinita a transforma em finita. Para fundamentar esta transformação partimos de um modelo simples inspirado pela "expansão do espaço económico", fundamentado com dados estatísticos de vários países recolhidos ao longo de várias décadas. Esta análise indica que é possível normalizar uma distribuição *heavy-tail* e também que é necessária uma análise extra na comparação de diferentes instantes de tempo.

Levamos este estudo um pouco mais longe para modelar também o efeito da inflação económica. A monitorização e controlo deste fenómeno é um dos pontos mais comuns em programas eleitorais, em declarações de missão de entidades internacionais ou em listagens de serviços de consultoras mundiais. Usando conceitos de teoria de grafos conseguimos expandir a análise inicial para o eixo do tempo. Este último passo da transformação permite aumentar os dados disponíveis que, tendo já sido normalizados para variância finita, servem para melhorar a estatística no estudo da Economia e dos agentes e sistemas que a compõe.

Palavras Chave: Economia, *Heavy-tail*, Teorema do Limite Central, Distribuição de Pareto, Lei de Potência, Redes Complexas, Valor do Dinheiro, Empresas Portuguesas.

Abstract

Two fundamental concepts in Finance are the **assets** and the **liabilities** of a certain entity. What I have (assets) and what I owe (liabilities) are concepts present within common sense and believed to be an indication of "economic health" of companies and individuals. Many have attempted to extract information and obtain predictive power from these economic values, but actually, "measuring the economy" is a challenge in and of itself. While it is easy to know the amount of currency in circulation, that is just a representation of the economy. What we truly want to measure are the trades of goods and services, along with the "work" (here we use the human concept, not the physics one) that created them in the first place. –On top of this we have the highly confidential and personal nature of the data in question.

A difficult point to overcome that has been downplayed in most literature is that the distributions of an economic variable (wealth, debt, income, etc.) are usually non-Gaussian. In fact, the variance of these distributions is usually infinite, which has earned them the classification of *heavy-tail* distributions in economic literature. With infinite variance the applicability of the Central Limit Theorem is highly impaired which means more data will not necessarily help. In fact, the value of statistical moments does not converge with the increase of statistical data.

This text sugests a transformation algorithm that, when applied to a distribution with non-finite variance will map it into one with finite variance. To justify the steps taken we will start from a simple model inspired by the expansion of the economic universe, and emboldened by statistical data from several countries over several decades. This analysis also indicates that it is indeed possible to normalize a *heavy-tail* distribution and an extra step is required when comparing different points in time.

We will take this model a bit further by looking the effects of inflation. Monitoring and controlling this phenomenon is a prevailing point in electoral programs, mission statements of international entities, service offered by world-wide consulting firms. Using ideas and methods from graph theory we are able to expand our initial analysis to the time axis. This will greatly increase the amount of data available to study the Economy and the systems that are part of it.

Keywords: Economy, Heavy-tail, Central Limit Theorem, Pareto Distribution, Power Law, Complex Networks, Value of Money, Portuguese Companies.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

No dia-a-dia estamos habituados a que aquilo que nos rodeia tenha uma escala característica: os adultos têm uma altura média (1,73 m[1]), o trânsito citadino tem uma velocidade máxima de circulação (50 km/h), um café tem preço típico (0,50 € a 1,00 €). Isto permite aceitar uma série de premissas para prever, por exemplo, o tempo que uma viagem vai demorar ou o nosso orçamento do mês. De facto, isto está tão enraizado no nosso raciocínio que nem questionamos quando é válido ou não aplicar este princípio.

Tomemos um exemplo simples: a altura média dos trabalhadores de um edifício de escritórios. No primeiro andar temos um escritório de advocacia em que todos terão uma altura próxima dos 1,7 m. No segundo andar temos um infantário: aqui a distribuição terá dois valores característicos (1,7 m para os adultos, 1,1 m para as crianças de 5 anos). O terceiro andar é uma associação de aposentados e em diante. Se o prédio tiver um número infinito de andares, o *teorema do limite central* (TLC) diz-nos que a distribuição assimptótica da média de alturas segue uma *distribuição normal*, sejam quais forem as distribuições de cada andar. A importância deste resultado não deve ser menosprezada: é isto que nos permite fazer estudos estatísticos onde assumimos que a grandeza em estudo tem uma média e um desvio padrão com significado relevante o que por sua vez permite, por exemplo, excluir *outliers* ou prever tendências. Quem decidir criar uma empresa de pronto-a-vestir pode estar confiante que ninguém precisará de calças com 3 m de altura ou que um casaco para adulto certamente terá mais de 15 cm de largura.

O TLC estabelece que a distribuição da agregação das amostras é *assintoticamente* Gaussiana. Ou seja, só com um número de amostras suficientemente alto é que obtemos esta distribuição e este "número suficiente" vai depender da distribuição original de valores sendo que por vezes o hipotético valor a que chegamos pode nem fazer sentido (por exemplo a média de alturas no infantário estará entre 1,1 m e 1,7 m sem que haja necessariamente um indivíduo com essa altura).

[1] fonte "Homens portugueses entre os que mais cresceram nos últimos cem anos" jornal Público, notícia de 26/julho/2016, acedida em 16/abril/2020.

Voltemos ao nosso edifício onde, em vez de estudar a altura, estudamos quanto é que cada pessoa gasta numa refeição. No infantário haverá uma preocupação com reduzir o custo das refeições. Os advogados do primeiro andar vão almoçar fora todos os dias e no último andar, na *penthouse*, uma celebridade come *wagyu* e caviar todos os dias. Se fizermos o mesmo tratamento estatístico a esta grandeza, haverá conclusões úteis a que podemos chegar?

Uma das formas de interpretar este problema é que o dinheiro não se mede todo da mesma forma: enquanto que uma criança do infantário irá vangloriar-se por ter uma goma a mais do que o colega do lado, um grupo de advogados poderá discutir os méritos de Mustang vs Jaguar. Visto de outra forma: ninguém tem o dobro da altura média do país, mas facilmente encontramos quem seja 10 ou 100 ou várias ordens de grandeza mais rico que a "média" (na verdade não existe média porque, como veremos, a distribuição de riqueza não tem momentos estatísticos definidos).

Uma manifestação deste problema é o conceito de distribuições *heavy-tail* cuja distribuição segue uma lei de potência em vez de uma Gaussiana. As distribuições com variância infinita contrariam o TLC.

O conceito de *heavy tail* foi primeiro observado em 1965 por Price na distribuição de citações de artigos científicos (aqui ainda não é usado este termo) [1]. O conceito foi expandido em 1999 por Barabási, que usou o termo *scale free* (livre de escala) para ligações na World Wide Web [2]. Este comportamento foi observado para várias redes sociais e biológicas, sendo que a certeza desta classificação (rede sem escala vs aleatória) ainda está em debate. Alguns exemplos incluem a participação de atores em filmes, relações presa/predador num dado bioma, chamadas telefónicas. Este conceito é expandido, com a devida bibliografia, no Capítulo 3.

O presente documento propõe uma métrica que, aplicada a dados com distribuição livre de escala, resulta numa distribuição quase Gaussiana e iremos medir a "Gaussianidade" dos dados usando a variância, curtose e simetria (mais especificamente, passaremos de variância infinita para finita). Uma das aplicações mais interessantes deste método, é a área de *Machine Learning*, onde o TLC é fundamental. Um dos principais objetivos desta área de conhecimento é extrair informação de uma grande quantidade de dados. No entanto, se estes dados incluírem medições com variação infinita, a incerteza aumenta com o tamanho da amostra.

A estrutura desta tese é a seguinte:

1. **Introdução**;

2. **Motivação:** motivação para este trabalho, incluindo um estudo estatístico sobre a distribuição de riquezas nas populações de diferentes nações, assim como a sua evolução ao longo de três décadas, usando dados disponibilizados pela *World Inequality Database* (`https://wid.world/`). Este capítulo mostra a dificuldade na caracterização estatística de grandezas económicas e a necessidade de algum tipo de normalização;

3. **A economia como rede complexa:** o capítulo anterior foca na análise de uma população para um

dado instante. No entanto, a economia é um sistema em constante evolução. Assim, para comparar populações em diferentes instantes de tempo será necessário um passo extra no nosso tratamento, sendo este motivado pela modelação da economia como grafo e o respetivo comportamento emergente;

4. **Modelo:** apresentação do modelo, que consiste num algoritmo de transformação de distribuições. Os capítulos **Motivação** e **A Economia como Rede Complexa**, apesar de aparentemente desconexos entre si, são ambos necessários para fundamentar o modelo como um todo;

5. **Resultados:** Aplicação do modelo a uma amostra de ativos e passivos de empresas portuguesas entre 2000 e 2010;

6. **Conclusões e trabalho futuro:** um breve resumo sobre o que foi apresentado e discutido neste documento.

Capítulo 2

Motivação e Evidências

O dinheiro (seja liquidez económica, rendimentos, riqueza, etc.) é um exemplo de um espaço de medida, em que a distância entre dois pontos é mais pequena quanto mais próximos da origem eles estiverem. A Figura 2.1 mostra uma representação pictórica deste fenómeno.

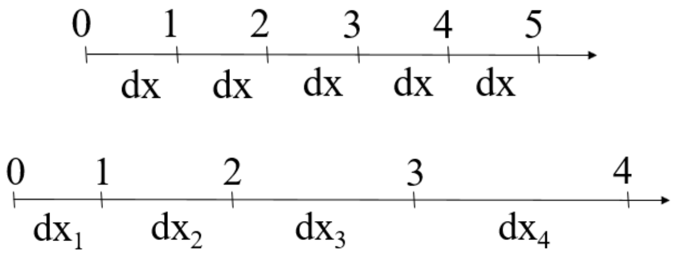

Figura 2.1: Num espaço linear (esquerda) a distância entre dois pontos é independente da posição desses mesmo pontos, o que não acontece da segunda imagem (direita).

O objetivo da nossa métrica será transformar a segunda figura na primeira, tornando a distância entre pontos independente das suas posições. Que relação poderá ser essa? A mais simples será a linear, dada por:

$$dx = \beta x \Leftrightarrow \beta = \frac{dx}{x} \Rightarrow \beta = \log(x_{\text{final}}) - \log(x_{\text{inicial}}). \tag{2.1}$$

Esta hipótese sugere que, olhando para a distribuição de distâncias de uma população a um ponto de referência, teremos algo com a forma de uma exponencial, muitas vezes identificada no contexto da economia.

2.1 A Economia vista por Físicos

Estudar a economia é um desafio com raízes no livro "*The Economics of Industry*" escrito por Mary e Alfred Marshall, publicado no século XIX. Um problema imediato: como medir *a Economia*? Se uma das definições é, seguindo o *Oxford English dictionary* "a área do conhecimento dedicada ao estudo da produção, consumo e transferência de riquezas" como é que definimos riqueza? As posses materiais de um individuo, ou a sua liquidez? O que já gastou ou o que está para ganhar? Mais: dada a natureza confidencial deste campo e a gigantesca tarefa que é registar toda a informação necessária para uma população, a simples *medição* torna-se um enorme desafio por si só.

No início da década de 90, um campo denominado Econofísica surgiu com a aplicação de métodos de Física Estatística a problemas de economia, mais especificamente a analise e previsão do mercado financeiro de ações e seus derivados tendo mais tarde evoluindo para outros campos. Uma grande motivação para explorar estes métodos foi reforçada com a enorme quantidade de informação que começou a estar disponível a partir da década de 80 com a adoção generalizada do uso de sistemas informáticos e da Internet para manter as bases de dados de empresas e governos. Logo no primórdio da Econofísica como ramo do conhecimento, foram "transferidos" conceitos da mecânica newtoniana e termodinâmica clássica [3], culminando no conceito de equilíbrio de *forças* (de procura e oferta) assim como uma tentativa de paralelismo com o conceito de energia livre [4]. Outros esforços incluem modelos de equilíbrio como o modelo Black-Scholes [5] para preços de derivados (esta linha de investigação explorada desde o início da década de 70) e a equação Fokker-Planck usada para modelar rendimentos [6, 7] . Um problema fundamental com esta técnica, ou qualquer técnica que use uma *master equation*, é a premissa assumida de conservação do espaço de amostras [8]. Mais especificamente, usar um sistema de *master equations* implica assumir que o sistema em estudo pode ser modelado pela combinação probabilística de um numero fixo de estados, com a probabilidade de transição entre estados dada por uma matriz. Por exemplo em Física Estatística, um *conjunto microcanónico* assume a conservação da energia total, o que significa que o número total de microestados é constante ao longo do tempo.

Por definição de sistema económico (cf. [9]), a "energia económica" cresce porque o "trabalho económico" de um agente não destrói o trabalho económico realizado anteriormente. Podemos argumentar que consome recursos, mas é o trabalho que dá valor a esses recursos. Assim, em vez de uma modelação macroscópica, uma outra possibilidade é uma análise microscópica: dando um conjunto de regras a cada elemento de um conjunto de agentes podemos simular um sistema de agentes e observar o comportamento emergente. Este é o princípio de funcionamento da modelação baseada em agentes (*agent-based modeling*).

Podemos sentir a tentação de definir o dinheiro como quantidade conservada: para além da omnipresença em todo o sistema económico, só bancos centrais podem imprimir moeda, o que significa que os podemos designar como fonte de energia externa e dizer que a economia de um dado país é um sistema fechado em equilíbrio macroscópico. Esta idealização de equilibro é muitas vezes usada em economia (cf. [9]), mas quase impossível de obter no mundo real: fenómenos como a entrada de moeda externa ou a cobrança de juros sobre crédito leva à criação de dinheiro, mesmo sem impressão de moeda nova pelos bancos centrais.

Um resultado comum na literatura, seja em estudos estatísticos [11], seja em simulações de agentes [12] é a emergência de uma distribuição com duas partes. Dando o exemplo da distribuição de rendimentos de uma população, obtemos um limite de rendimentos baixos próximo de um decaimento exponencial e um limite de rendimentos altos modelável pela distribuição de Pareto. Uma explicação sugerida por [13] para esta separação é que a riqueza da população mais pobre é composta por dinheiro, enquanto que a mais rica envolve outro tipo de bens como ações, investimentos, etc. O autor vai mais longe e sugere que este regime de riquezas superiores é um efeito especialmente fora de equilíbrio, requerendo constante expansão económica e que desaparece em sistemas fechados. Na próxima secção esta bipartição será evidenciada em dados estatísticos de vários países.

2.2 Distribuição de Pareto e WID

Em 1909, Vilfredo Pareto observou a forma da distribuição de riquezas e afirmou que seguiam uma distribuição semelhante a uma lei de potência que mais tarde se chamou **distribuição de Pareto** [14]. Hoje em dia, esta lei é observada em vários fenómenos como ligações entre *websites* [15, 16], tamanho de cidades [17, 18], severidade de desastres naturais [19–21], etc.

Para esta distribuição, a função de distribuição acumulada é dada por:

$$F_X(x) = \Pr(X \leq x) = \begin{cases} 1 - \left(\frac{x_\mathrm{m}}{x}\right)^\alpha & x \geq x_\mathrm{m} \\ 0 & x < x_\mathrm{m} \end{cases}, \tag{2.2}$$

a função densidade por

$$f_X(x) = \Pr(X = x) \begin{cases} \frac{\alpha x_\mathrm{m}^\alpha}{x^{\alpha+1}} & x \geq x_\mathrm{m} \\ 0 & x < x_\mathrm{m} \end{cases}, \tag{2.3}$$

e a função de sobrevivência, que é a complementar de (2.2), por

$$\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_\mathrm{m}}{x}\right)^\alpha & x \geq x_\mathrm{m} \\ 1 & x < x_\mathrm{m} \end{cases}. \tag{2.4}$$

Note-se que a *Distribuição* de Pareto não deve ser confundida com o *Princípio* de Pareto que diz que 80% da riqueza é detida por 20% da população, algo que se verifica apenas para o valor especifico de $\alpha = \log_4 5 \approx 1.16$. Alguns exemplos de representação gráfica destas funções podem ser encontradas na Figura 2.2.

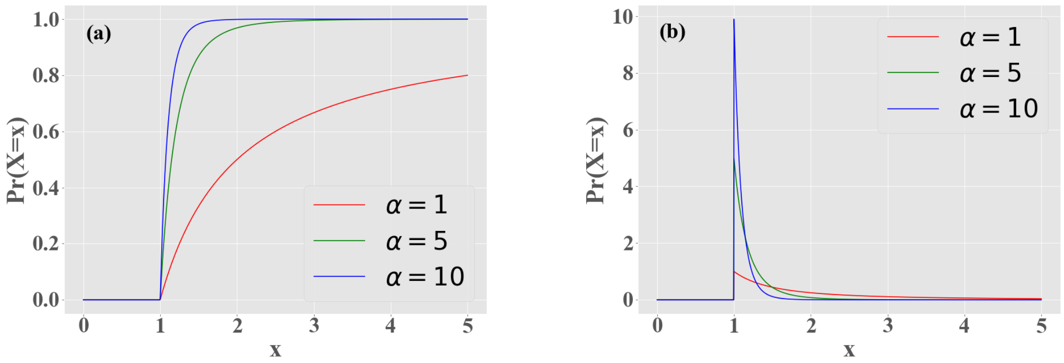

Figura 2.2: Gráficos exemplificativos da distribuição de Pareto para uma variável genérica X, função distribuição acumulada (a) e função densidade de probabilidade (b). No infinito, (a) tende para função degrau e (b) para o Delta de Dirac.

World Inequality Database (WID) (inicialmente conhecido como *The World Top Incomes Database*) é um projeto iniciado pelo economista Thomas Piketty em 2011 com o objetivo de catalogar a distribuição de rendimentos e riquezas por todo o mundo, compilando trabalhos de diversas nações feitos por diversos autores e investigadores [22]. Uma das ferramentas desenvolvidas por este grupo é o GPinter (de *generalized Pareto interpolation*, cf. [23]), que usa uma distribuição de Pareto generalizada para interpolar a distribuição de riquezas numa população a partir de informação tabelada (vd. `https://data.worldbank.org/`), muitas vezes disponibilizada por agências governamentais ou institutos estatísticos. Atualmente, o *website* `https://wid.world` mantém um repositório de tabelas com dados sobre riqueza, salários, e outras variáveis económicas de diversas nações, compilados por diferentes investigadores. Usaremos estes dados para demonstrar a aplicabilidade da distribuição de Pareto a dados económicos assim como evidenciar uma conclusão fundamental.

Alguns pontos relevantes sobre os dados disponibilizados pela WID:

- A grandeza económica em estudo neste documento é *pre-tax national income* (rendimento nacional pré-impostos). Esta é o somatório dos rendimentos (*income flows* no original) para cada indivíduo que é dono de fatores de produção, trabalho, capital, antes da atuação do sistema de impostos mas depois da operação do sistema de pensões, desemprego e outros sistemas de segurança social;

- *National income* mede o rendimento de um país que está disponível para os seus habitantes. É igual ao produto interno bruto (GDP, valor total dos bens e serviços produzidos no território nacional num dado ano) menos o capital fixo usado em processos de produção (e.g. reparação e substituição de equipamento, manutenção de estradas) mais o resultado líquido de rendimentos estrangeiros obtidos por residentes no resto do mundo;

- Para comprar diferentes países e anos, a WID usa o conceito de *purchasing power parity* (PPP). Segundo este conceito, duas moedas estão em equilíbrio quando um cabaz de produtos tem o mesmo preço em ambos os países (considerando taxas de câmbio). Os dados apresentados neste trabalho foram convertidos, pela própria WID, para PPP euro 2016 não estando disponíveis noutra forma;

- Alemanha e Reino Unido não disponibilizam as distribuições de rendimentos, apenas os quantis mais altos (*top income shares*). A restante distribuição é extrapolada com base nesses valores e nos valores de países vizinhos.

Para informações mais detalhadas sobre a metodologia deste trabalho recomendamos [24] e [25] assim como a extensa bibliografia produzida pelo grupo [1].

WID, assim como diferentes publicações [11], modelam a distribuição de rendimentos ou de riquezas com uma lei de Pareto para valores altos (i.e. *upper tail*) e o resto com uma lei de potência mais geral, ou até mesmo uma distribuição de Pareto com um coeficiente diferente ou variável (i.e. uma distribuição de Pareto generalizada).

2.3 Modelação da distribuição de riquezas

Para demonstrar a validade do uso de uma distribuição de Pareto na modelação de variáveis económicas, usamos os dados obtidos da WID para estimar o expoente da distribuição de graus para diversos países em diversos anos.

A título de exemplo, a Figura 2.3 mostra a distribuição de *pre-tax national income* em França para o ano 2000, onde a linha a roxo representa a distribuição de Pareto para $\alpha = -2,7$. Podemos desde já observar um claro ponto de inflexão (em $x \approx 9,5$ o que corresponde a um salário anual de 13359,7 €). O exemplo apresentado ilustra a dificuldade já expressada em modelar toda a população com uma única distribuição, levando vários autores a separar os rendimentos mais altos dos mais baixos (cf. [14, 26, 27]). No ano 2000, o salário mínimo nacional anual em França foi 12796,8 €[2], o que sugere ser uma possível explicação para este ponto de inflexão.

Um entendimento empírico deste corte é que todos os rendimentos abaixo do salário mínimo nacional são rendimentos passados (pensões e subsídios de desemprego são proporcionais ao salário antigo) podendo ser considerados como riqueza já gerada (este ponto é expandido na secção 3.2). Com um *cutt-off* para rendimentos abaixo do salário mínimo nacional obtemos um expoente de 2,7. Este processo foi executado para os anos 1980, 1990, 2000, 2010 e para Alemanha, Brasil, Estados Unidos da América, França, Reino Unido, Rússia. A Figura 2.4 mostra a variação do ajuste para cada país.

Fazemos notar que os valores usados estão indexados ao Euro em 2016, por isso as retas não apresentam translações segundo o eixo horizontal que seriam causadas pela inflação. Para além de todos os

[1] https://wid.world/methodology/#library-general, acedido em 29/Set/2019.
[2] https://countryeconomy.com/national-minimum-wage/france, acedido em 19/Ago/2019.

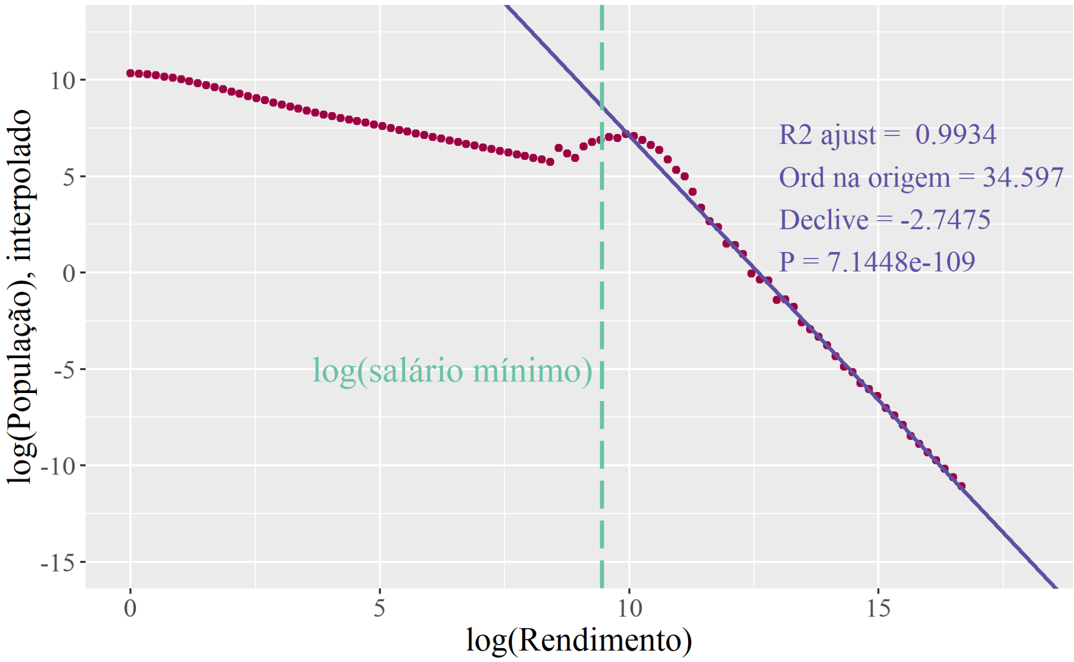

Figura 2.3: Histograma de *pre-tax national income* em França, para o ano 2000. O ajuste é feito desde a linha de salário mínimo nacional anual. Nesta figura podemos observar os dois regimes da distribuição: abaixo do salário mínimo temos uma distribuição quase uniforme, acima temos o comportamento log-log esperado com um declive bem definido.

países apresentarem esta mesma forma geral, vemos que ela se mantém ao longo de décadas, mesmo atravessando alterações geopolíticas como a queda da União Soviética (1989), a mudança de moeda (1999 para França e Alemanha) e a bolha imobiliária e subsequente crise do *subprime* de 2007.

Esta análise foi desenvolvida numa tentativa de transformar a distribuição de uma grandeza económica de forma a que esta seja modelável. Este é o ponto de partida do modelo apresentado nesta tese, que será explicado no capítulo 4. Não fazendo parte do algoritmo final, é uma parte integrante na fundamentação do modelo como um todo e necessária aquando a aplicação do mesmo a outros conceitos.

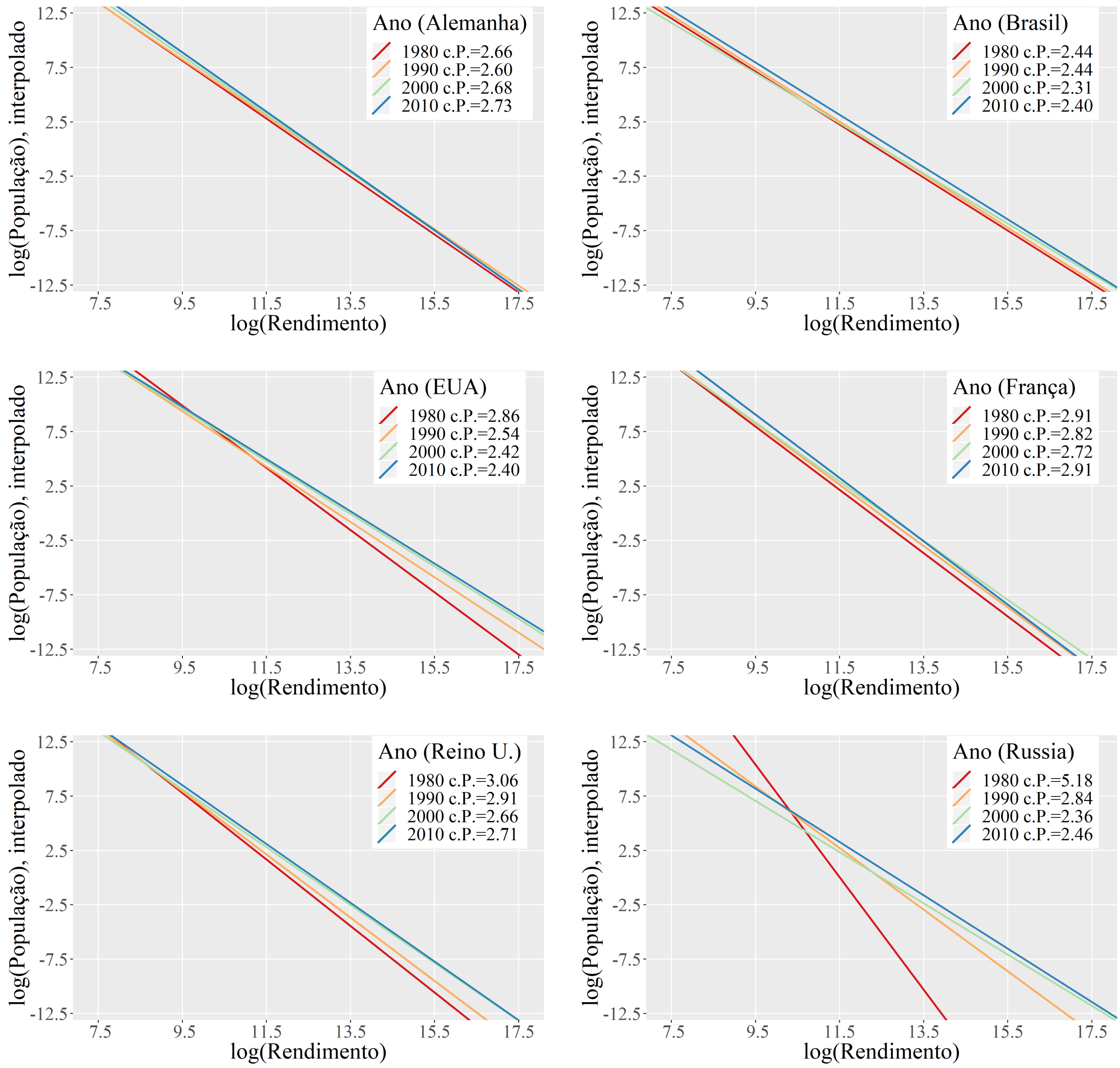

Figura 2.4: Evolução da distribuição de *pre-tax national income* para os anos 1980, 1990, 2000, 2010. Todos os coeficientes de Pareto encontram-se entre 2 e 3, à exceção de Rússia e Reino Unido em 1980. O comportamento observado na Figura 2.3 é comum não só em diversos países mas também ao longo de três décadas.

Capítulo 3

A Economia como Rede Complexa

Quem começa o seu dia com uma bica, fá-lo acompanhado de uma troca económica: o cliente dá o seu dinheiro em troca do produto (café). Mais tarde, quando precisa de comprar um bloco de notas e uma caneta, mais uma troca. Depois tem que pagar a fatura da luz: desta feita, o dinheiro do cliente é retribuído com um serviço. Podemos representar cada uma destas entidades como um ponto, e cada troca com uma linha que une esses pontos (Figura 3.1a); ao fim do dia teremos algo parecido à Figura 3.1b. Mas uma pessoa não é o centro do universo: todos os pontos (a que chamaremos agentes económicos) fazem trocas entre si, o que significa esta rede de ligações económicas será algo mais parecido à Figura 3.1c.

Todas as figuras apresentadas são chamadas *grafos*, sendo a teoria dos grafos um ramo da matemática dedicado a estudar as relações entre elementos de um conjunto. Assim, os grafos são formados por nós e ligações. Quando um grafo apresenta algum tipo de comportamento emergente, é usual usar o nome *rede complexa*. Este conceito começou a ser estudado em 1735 quando Euler usou um esquema semelhante para resolver o então chamado "Problema das Sete Pontes de Königsberg" [28]. Este campo do conhecimento foi ganhando ímpeto quando se percebeu que, com alguma abstração, se podia aplicar a diversas áreas (p.e. a WWW é formada por *websites* e as hiperligações entre eles; as relações sociais por pessoas e as ligações entre elas; os ecossistemas por espécies e relações presa/predador). A aplicabilidade desta descrição aumenta quando se modifica as funcionalidade da rede usando:

- Redes dirigidas: separar as direções de cada ligação (por exemplo, separar o produto/serviço do pagamento);
- Redes bipartidas: diferentes tipos de pontos (por exemplo, atores de cinema e os filmes onde participam) sendo que um nó só se pode ligar a nós do outro tipo.

Uma rede pode chegar rapidamente aos milhões ou biliões de nós e ligações. Como é que podemos caracterizar a forma de uma rede se não a podemos visualizar? Este é um dos principais pontos de estudo desta área do conhecimento e para uma introdução, remetemos o leitor para as fontes [29–31] que serão amplamente citadas nesta secção.

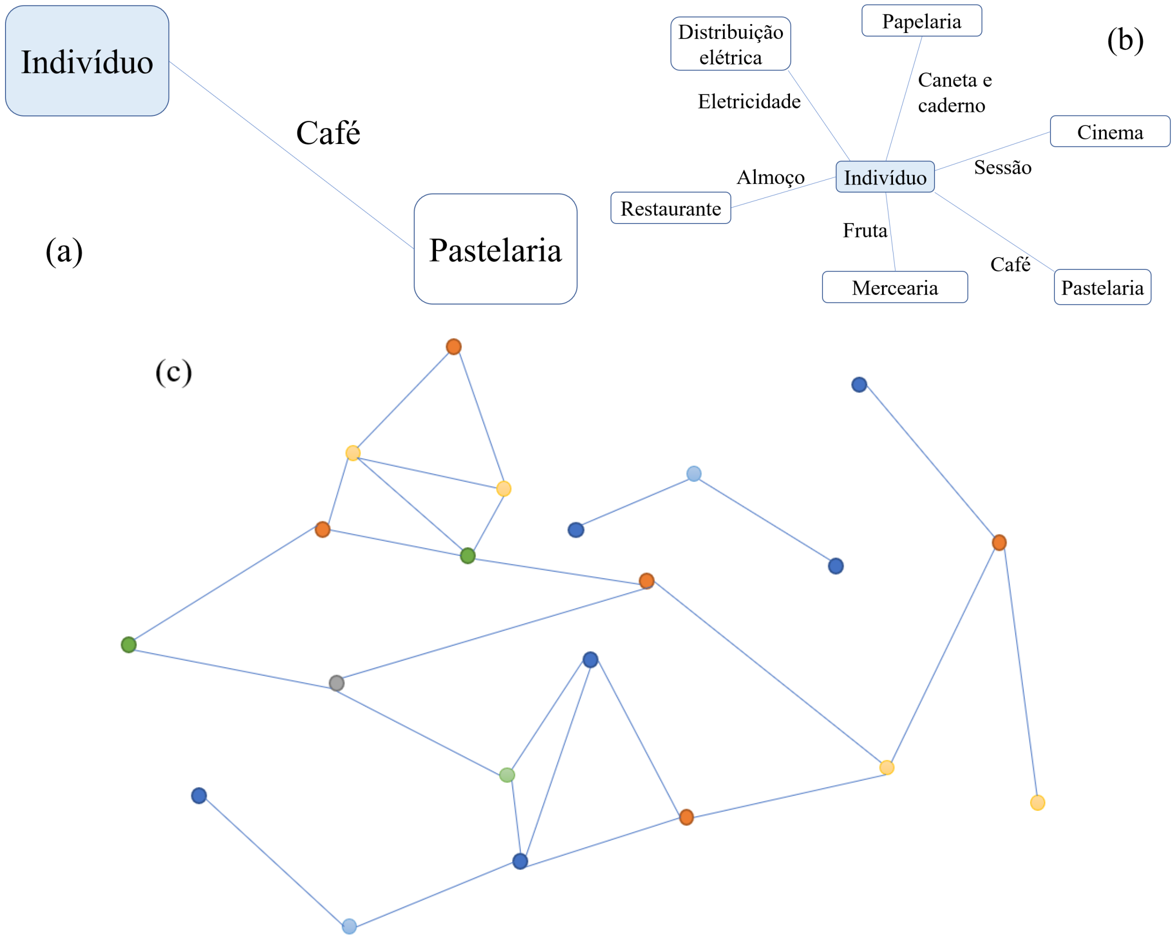

Figura 3.1: Representação num grafo de diferentes trocas económicas. Em a) vemos uma troca económica entre dois agentes, em b) um conjuntos de trocas que um mesmo agente pode fazer e em c) uma representação das trocas entre vários agentes.

Um dos conceitos mais simples no estudo de grafos é o **grau do nó**: o número de ligações de um dado nó (Figura 3.2). Podemos pensar que o "caos" da vida levaria a que as ligações que formamos em torno de nós seriam "aleatórias" e sem padrão (estas hipotéticas redes são designadas por *random networks*, redes aleatórias), o que levaria a que a distribuição Gaussiana de graus (muitos nós perto da média e poucos nos extremos). Em vez disso, quando fazemos a estatística de redes reais e estudamos a distribuição de graus, obtemos em muitos casos uma lei de potência da forma $x^{-\alpha}, \alpha > 0$ com expoente constante. A estas redes chamamos *scale free* (livre de escala). Uma nota sobre o que queremos dizer com expoente constante: modelando diferentes sistemas como uma rede complexa obtemos para cada tipo de sistema um expoente característico ($\approx 2,3$ para atores de cinema, $\approx 2,5$ para redes informáticas, $\approx 2,1$ para telefonemas, etc [29]).

Figura 3.2: Exemplo de uma rede complexa com seis nós e cinco ligações. Um dos nós tem grau 4 (quatro), dois nós têm grau 2 (dois), dois nós grau 1 (um) e um nó tem grau 0 (zero).

Assim, parte do trabalho de investigação nesta área foi dedicado à modelação do fenómeno que causa esta acentuada diferença nas distribuições. Um dos conceitos com mais sucesso chama-se "*cumulative advantage*" ou "*preferential attachment*" (a maioria dos estudos consideram *preferential attachment* linear podendo este ser superlinear ou sublinear [29], mas esse estudo fica fora do âmbito desta tese). Este coeficiente característico das redes livres de escala (normalmente representado na bibliografia por α) está normalmente entre 2 e 3. Barabási e Albert (cf. [29–31]) dedicaram algum esforço a estudar estes extremos, chegando a dois mecanismos para construir redes que levam à distribuição pretendida:

- **Mecanismo A:** rede que cresce sem ligação preferencial. Começando com m_0 nós, a cada iteração adicionamos um nó que se ligará a cada outro nó já existente com probabilidade uniforme; este novo ponto na rede terá $m(\leq m_0)$ ligações.

- **Mecanismo B:** a rede evolui apenas por ligações. Começando com N nós sem ligações, a cada iteração um nó aleatório é ligado a outro nó i com probabilidade $P(k_i) = k_i / \sum_j k_j$, sendo que o número de nós se mantém constante.

A teoria prevê que o mecanismo A leva a uma distribuição de graus exponencial (o que significa que não é livre de escala). Simulações numéricas mostram, para o mecanismo B, a existência de um limite inferior a partir do qual se mantém uma lei de potência, mas esta tende a transformar-se em Gaussiana depois do transiente inicial. Dado que nenhum dos mecanismos, individualmente, leva a uma rede *scale free*, podemos assumir que ambos são necessários em alguma proporção.

Distribuições em lei de potência são consistentemente observadas em variáveis económicas e em redes complexas de fenómenos sociais. Esta ligação indica que a área de conhecimento explorada neste capítulo pode ser útil na construção do modelo.

3.1 As trocas económicas

O corpo de conhecimento que existe sobre redes complexas é imenso. Conhecendo a distribuição de graus e fazendo a correspondência certa dos conceitos de redes com as observações, as conclusões são aplicáveis a qualquer área (uma das provas disto é que o estudo de resiliência de um grafo tem aplicações práticas na segurança e manutenção de redes informáticas e em estratégias de vacinação [29–31]). Mas então, quão válido é fazer a correspondência para o sistema económico de nó/agente e aresta/troca?

A economia é formada pelos agentes (representados pela sua riqueza, pelo seu poder de compra, pela sua dívida, etc.) e pelas trocas entre eles (compra, venda, crédito, etc.). Esta abstração permite que nos foquemos numa medição específica em vez de tentar analisar todo o sistema. Por exemplo, podemos olhar para os dados de uma instituição de crédito com uma quantidade significativa de clientes para estudar a distribuição de dívida; podemos estudar os registos de um banco para conhecer a distribuição de ativos e passivos. Mesmo não conhecendo mais informação sobre os agentes individuais (o que ajuda imenso com questões de privacidade e sigilo), uma lista de valores numéricos já pode dar uma ideia de como essa grandeza económica se comporta na população.

Independentemente da forma física, legal ou social do conteúdo dessas trocas, elas representam uma quantidade comum que é acumulada por cada agente. Aqui, quando dizemos "acumulada" talvez seja mais fácil perceber como um conceito "contabilístico". Quando um agricultor produz e vende fruta a uma mercearia, o merceeiro irá vender aos seus clientes que por sua vez a irão consumir. Aqui, a fruta não é realmente acumulada, mas a quantidade cumulativa de fruta comprada e vendida vai aumentando. Enquanto isto acontece, haverá fornecedores que recebem mais encomenda do que outros (seja por sorte, seja por competência, etc.) e que poderá reinvestir parte do dinheiro recebido na expansão do seu negócio (mais área de cultivo, novas máquinas, melhores adubos). Com o aumentar da população a procura por fruta vai aumentar, o que vai levar superfícies comerciais a procurarem os fornecedores de fruta que conseguem oferecer em maior quantidade, levando a que o maiores cresçam mais. Este último ponto é o que chamamos *preferential attachment*, que pode ter origem aleatória (por sorte inicial, um produtor é mais escolhido do que o outro) ou por outros motivos (se um produtor é mais competente ou vende mais barato, aqui serão as forças de mercado a definir os alvos de *attachment*).

Com o crescer dos agentes envolvidos em trocas, toda a economia cresce ao longo do tempo.

3.2 Definir o dinheiro

Como é que definimos o valor monetário de um bem ou serviço? Ao longo da história houve várias hipóteses como por exemplo a soma do custo de produção (incluindo matéria prima e mão de obra) ou *o que o mercado está disposto a pagar* (o que é especialmente notório, por exemplo, em obras de arte).

Um conceito que surge nesta discussão é o de escassez (*scarcity* em inglês). Num contexto económico este termo refere-se não à falta de recursos mas serve para salientar que os recursos são finitos em contraste às necessidades que podem ser, em teoria, infinitas (isto porque os recursos são passados já foram produzidos, e as necessidades são futuras, são cumulativamente infinitas). Podemos então pensar na utilidade ou raridade de um recurso: se algo é muito útil ou indispensável ao ser humano ou à sociedade em geral, esse recurso será mais cobiçado. Se isto fosse verdade, um copo de água seria mais valioso que um diamante. E raridade? Na industria alimentícia ovos podres são mais raros que ovos bons, e estes é que têm valor, não os primeiros.

Podemos ir a um ponto mais profundo: porque é que existe economia? É comum dizer que o ser humano é um animal social: vivemos juntos e colaboramos. Apesar de não ser impossível, é bastante difícil que um individuo consiga satisfazer todas as suas necessidades individualmente. A necessidade de colaboração leva a trocas e com um número suficiente de trocas, um sistema irá emergir a que damos o nome de *mercado*.

Imaginemos um agricultor que tem um hectare de solo arável que usa para produzir fruta e legumes. O seu objetivo é adquirir uma casa. Matematicamente haverá uma certa quantidade de frutas e legumes (x) que alguém estará disposto a trocar por uma casa (ou seja, podemos pensar em x como uma quantidade abstrata que vale uma casa ou um certo volume/peso de legumes). Há um limite para a quantidade que o agricultor consegue produzir, e ele não pode simplesmente acumular o que produz até ter uma quantidade suficiente para trocar por uma casa. Em vez disso, vende o que produz em troca de dinheiro, e acumula esse dinheiro até conseguir comprar a casa. Assim, dinheiro que o agricultor tem representa todo o trabalho que foi feito no passado. Note-se que o dinheiro em si não é verdadeiramente o objeto da troca: ele representa uma "suspensão" do trabalho já executado que mais tarde será trocado pelo trabalho de outra entidade: eu acumulo o valor do meu trabalho sob a forma de dinheiro, e uso-o para adquirir alimentos, pagar a conta da eletricidade, outros bens e serviços. Podemos então pensar no dinheiro como um referencial externo ao sistema: serve apenas como medida para o sistema económico de trocas de bens e serviços.

Aprofundemos um pouco mais o exemplo: o agricultor passa 10 semanas a cultivar uma certa quantidade de batatas que no final é vendida por 1000 €. Este dinheiro, que representa o valor do seu trabalho, é então guardado debaixo do colchão. De seguida passa 20 semanas a cultivar couves, e recebe 2000 €. Estes 3000 € que estão debaixo do colchão é todo igual? Pensemos na natureza material de uma nota de 20 €: uma determinada entidade (normalmente, um banco central) olha para todo o trabalho feito até aquele momento e imprime uma quantidade de notas que acredite que represente esse mesmo trabalho. Nesse instante, 1€ representa x unidades de trabalho (uma certa quantidade de batatas, ou de horas de investigação, ou de turnos de supermercado). Nesse instante, o observador Banco Central vê esta quantidade e as notas impressas e as moedas cunhadas representam essa quantidade. Como o passar do tempo, mais trabalho é gerado, mas o valor escrito numa nota de 10 € continua a ser 10 €, o que significa que 1 € passa a valer αx unidades de trabalho. Então, debaixo do colchão estão $(1000 + \alpha 2000) \times x$ unidades de trabalho. Note-se que isto significa que o dinheiro desvalorizou: quando o agricultor tem que comprar um sacho novo, a quantidade de metal e madeira e tempo de fabricação é o mesmo de há um ano atrás, mas serão "precisos mais euros" para o pagar; visto de outra forma podemos dizer que enquanto se usar um referencial fixo, “um joule” de trabalho no passado é mais valioso que “outro joule” no presente.

Como podemos estudar esta evolução? Pensemos então que depois da primeira leva de batatas, o agricultor investe os 1000 € em mais área de cultivo, melhor adubo, instrumentos modernizados. A segunda leva (assumindo que não houve nenhum desastre) será maior. Se a primeira leva envolve x unidades de trabalho, a segunda cresceu dx. Olhando para todos os camponeses, pastores, fabricantes, empreiteiros, etc. vemos que a relação entre os recursos que existiam e o que foi produzido:

$$\frac{1}{x}\frac{dx}{dt} = \beta. \tag{3.1}$$

Esta equação, vista para um dado instante, já surgiu aquando a nossa tentativa de linearizar um espaço (equação 2.1). Isto significa que, visto do instante em que o nosso agente *agricultor* recebe os 1000 € (x_0) pela primeira encomenda de batatas (t_0), a sua riqueza (sob a forma de nova dose de batatas a crescer mais equipamento mais o que ainda estiver líquido) cresce segundo

$$x(t) = x_0 e^{\beta t}. \tag{3.2}$$

Uma nota importante: dado que β é dado pela média de todas as entidades envolvidas pode ser usado para todos os agentes que formam o sistema. Ou seja, para as entidades a e b

$$\frac{x_a(t)}{x_a(t)} = \frac{x_a(0)(1+\beta)}{x_b(0)(1+\beta)} = \frac{x_a(0)}{x_b(0)}, \tag{3.3}$$

o que significa que o problema de desvalorização do dinheiro acontece porque usamos uma observação estacionária no tempo, se as medidas fossem feitas em unidades de trabalho de uma qualquer entidade de referência que evolua com o sistema, este problema não se apresentaria. Esta observação será especialmente útil na definição de observado comóvel do sistema em estudo.

Capítulo 4

Modelo

Pretendemos então construir um modelo de sistema económico que nos permita usar as ferramentas matemáticas necessárias para normalizar a população. Definimos os seguintes conceitos:

1. **Agente económico** é uma entidade que produz (um produto, um serviço, uma riqueza) com uma certa relação à quantidade de recursos que lhe é alocada;

2. **Sistema económico** é um conjunto de agentes económicos grande o suficiente para ser considerado infinito em termos práticos que interagem entre si;

3. **Riqueza (de um agente)** é a quantidade de recursos alocados a um determinado agente (representada por x), que resulta da acumulação de trocas já explicada;

4. **Produto (de um agente)** é o aumento dos recursos alocados a um agente aquando de uma troca (representada por dx).

Reiteramos que a importância do conceito de Produto, mais do que a sua natureza específica, é estabelecer uma relação entre os recursos económicos alocados a um agente e a contribuição que esse mesmo agente faz para o sistema. Mais uma vez observamos uma diferença fundamental entre a Física e a Economia: a primeira apoia-se em leis de conservação, na segunda, os conceitos como riqueza e dívida aumentam proporcionalmente ao seu tamanho o que faz com que o próprio sistema aumente, não havendo conservação. Uma pessoa trabalha e o seu esforço produz riqueza que leva à expansão do universo económico.

Se um agente económico produz (dx) em proporção ao que já tem (x), então podemos escrever que

$$dx = \beta x \Leftrightarrow \beta = \frac{dx}{x} = d \log x. \tag{4.1}$$

Nesta equação β é a nossa concretização matemática de **agente económico** ou mais concretamente, é a relação entre os recursos alocados a um agente e os que ele produz. Mais importante do que a sua definição exata é o facto de que nos leva a um processo multiplicativo ${}^{dx}/_{x}$.

Cada interação entre agentes, inclusivamente de um agente com ele próprio, causa um crescimento de todos os que participam nessa interação. Isto é conhecido como *processo multiplicativo* e é sabido que processos multiplicativos, a que Piketty [27] chama *choques multiplicativos aleatórios*, são um mecanismo fundamental no aparecimento de distribuições que seguem uma lei de potência. De facto, todos os processos multiplicativos levam a uma distribuição com a forma de uma lei de potência (ver a revisão feita por Sornette em [32, 33]).

Estudar o futuro Quando tentamos prever a evolução de um sistema, um conceito que surge imediatamente é *probabilidade*. Em 1933, Andrey Kolmogorov estabeleceu três axiomas que não só uniam as diversas interpretações de probabilidade (nomeadamente, clássica e frequencista) como também estabeleciam o formalismo matemático da probabilidade. Estes axiomas são dados por

$$P(\epsilon) \in \mathbb{R}, P(\epsilon) \geq 0 \qquad \forall \epsilon \in \Sigma, \tag{4.2}$$

$$P(\Omega) = 1, \tag{4.3}$$

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i). \tag{4.4}$$

Assim, para um evento ϵ temos um espaço de probabilidade definido por $(\Omega, \Sigma, \mathrm{P})$ onde Ω é o espaço de amostras, Σ o espaço de eventos e P a probabilidade de ϵ.

Uma das afirmações feitas nesta discussão é que a Economia é um sistema em constante expansão: um agente a que sejam alocados recursos gera riqueza ("dinheiro gera dinheiro"). Assim, normalizando uma variável económica (p.e. o número de transações numa população, o valor da dívida pessoal) num determinado instante t_0, com a expansão da economia (por crescimento dos negócios, por inflação monetária) esta normalização deixa de ser válida para $t > t_0$. Concluímos que a equação (4.3) não é verificada e por isso a medida de probabilidade num instante não é uma medida de de probabilidade no instante seguinte.

Tendo por base o trabalho de [34] vamos assumir que a distribuição *power law* é um bom modelo para uma variável aleatória que descreve um aspeto económico de uma população (aqui podemos ler rendimentos, dívida, riqueza, etc). Se não podemos usar o conceito de probabilidade para extrapolar o comportamento deste tipo de variáveis (ou seja, prever o futuro), será que a definição da distribuição em si é robusta o suficiente para prever a evolução dessa mesma população? Uma limitação já observada neste documento (Capítulo **Motivação**) é que a aplicação de uma lei de potência a fenómenos observados implica sempre um intervalo de validade. No caso da WID, a lei de potência é usada para o intervalo de rendimentos mais alto, com o limite inferior dado empiricamente por uma transição para decaimento exponencial, e no limite superior pelo agente mais rico ou pelo total de riqueza presente no sistema. Para além disso, dados que sigam uma distribuição com esta forma só tem momentos finitos para certos intervalos de expoentes, ou seja, para $x^{-\alpha}$ com $2 < \alpha < 3$, todos os momentos de segunda ordem e superior divergem [35, 36]), o que impossibilita modelar e fazer previsões. Isto reforça a nossa suposição de que a modelação de um ou mais aspetos económicos envolvem uma análise mais sofisticada do que a simples aplicação direta de um conceito individual.

4.1 Descrição do Problema

Até aqui discutiu-se que *a distribuição de uma variável económica numa população tende a seguir uma lei de potência*. E tal como já foi mencionado, se esta lei de potência tiver um expoente entre 2 e 3 (como aconteceu nos dados retirados da WID) os momentos estatísticos não estão definidos. O problema pode então ser resumido de uma forma muito simples: com variância infinita, qualquer interpolação ou extrapolação não é útil, qualquer "previsão" que façamos não é mais nem menos válida do que qualquer outra. Este problema é tão comum na área financeira que este tipo de distribuições é chamado *heavy tail*, referindo-se ao facto de apresentarem valores extremos com mais frequência do que seria expectável numa distribuição Gaussiana (note-se que esta definição não é completa, mas serve de indicação do problema mais comum que resulta destas distribuições).

Recordamos três pontos fundamentais já introduzidos:

1. Um sistema económico pode ser modelado por uma rede complexa;
2. Distribuições em lei de potência são difíceis de modelar;
3. O conceito de rede complexa por si só não é suficiente para um estudo rigoroso os fenómenos presentes.

Uma rede complexa é formada pelos nós e pelas suas ligações. Não há "espaço circundante" o que significa que toda a geometria depende do tempo visto que o espaço económico é formado pelos agentes que nele participam, evoluindo com eles (ver Figura 4.1).

O espaço económico é formado pelos agentes que nele participam, e uma das formas para medir este espaço é estudando as chamadas grandezas económicas: ativos, passivos, rendimentos, riquezas, etc. Com interação, os agentes crescem e o espaço cresce com eles e as "fronteiras" desse mesmo espaço são definidas pelos agentes nos extremos: em todo o mundo existe uma entidade com o valor da grandeza económica mais alta, e acima desse valor o espaço económico não existe. Ao mesmo tempo existe uma zona vazia, um limite inferior abaixo do qual não existem agentes e por isso não existe espaço económico (ninguém recebe 3 centavos de salário). Isto pode ser interpretado usando um conceito de zona de "economia ativa" que vai expandindo, com a inflação e outros fatores e que pode ser representada, novamente, pela Figura 4.1.

É isto que nos leva a invocar os axiomas de Kolmogorov: quando, num instante, estabelecemos um espaço de probabilidade, no instante seguinte o universo é diferente: riquezas aumentam ou diminuem, agentes são criados ou desaparecem, entre outras mudanças, fazendo com que o axioma da unidade de

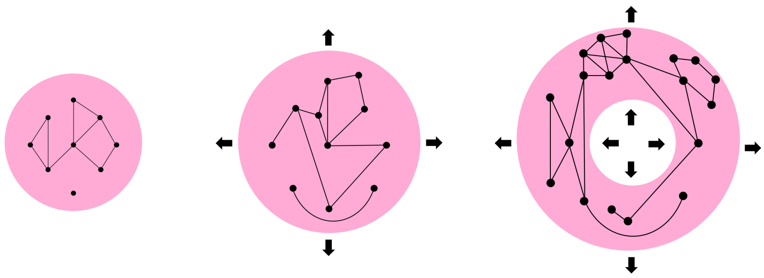

Figura 4.1: Representação pictórica da expansão do espaço económico. Podemos pensar que um agente cresce ("fica mais rico" ou "acumula dívida") à medida que que se afasta do centro e que um vazio ("zona morta") que se forma abaixo de um certo limite inferior já que os agentes não são criados do zero. Note-se que ao contrário de outros sistemas físicos, o espaço circundante (zona rosa) não existe: o espaço económico é formado pelos agentes (pontos pretos) e pelas ligações (linhas pretas) entre os mesmos.

medida não seja respeitado. Este aspeto, em conjunto com o facto dos momentos infinitos das distribuições observadas em sistemas económicos invalida qualquer previsão que possa ser feita.

Agente de referência O modelo proposto assume a existência de um observador comóvel para quem este espaço económico é uniforme. Dito de outra forma, este observador verá os restantes agentes económicos uniformemente distribuídos em torno dele. Para um determinado instante procuramos então uma normalização que transforme a distribuição da variável económica numa distribuição Gaussiana ou aproximadamente Gaussiana (usaremos a equação 4.1).

O vazio económico O último passo do algoritmo proposto envolve um fator de normalização entre diferentes instantes. A questão que se põe é: tendo uma população normalizada para um dado instante de tempo, como podemos comparar dois agentes (e, por extensão, duas populações) em instantes diferentes?

Voltando ao nosso diagrama empírico (Figura 4.1) vemos que a expansão do *donut* é acompanhada pela expansão do vazio no centro. O limite desta zona vazia é o agente mínimo: aquele (ou aqueles) cujo valor da grandeza em estudo é o menor da população, devidamente normalizado; o modelo apresentado baseia-se na posição do agente mínimo para definir a evolução do vazio que se cria na economia (o limite interior do *donut*). Este conceito é concretizado aquando a definição do fator de evolução (que designamos por fator c) que relaciona a variação do "tamanho" do centro do *donut* e que, em conjunto com a descrição da população em si, permite uma descrição grosseira da evolução económica. Um valor nulo deste fator significa que o "vazio económico" se manteve igual, um valor positivo implica uma expansão e um negativo uma contração. Por si, isto não significa que a economia como um todo tenha expandido ou contraído, significa que a "linha de água" abaixo da qual uma agente desaparece (i.e. falências) sobe ou desce. Numa economia dita *forte* este valor deve manter-se positivo de ano para ano sendo que um valor negativo pode significar uma recessão.

Note-se que este cálculo apresenta um desafio prático: o agente mínimo ideal não existe. No mundo real as empresas vão à falência ou são adquiridas ou possivelmente têm uma brilhante ideia de negócio que as catapulta para o topo da Forbes Global 2000[1]. Para mitigar esta situação, usa-se a média dos agentes mais baixos para simular este agente.

4.2 Descrição do algoritmo

Se queremos normalizar uma variável económica existem apenas dois conceitos que podemos usar: agente ou ligação (o nosso entendimento de economia, para esta tese, é formado apenas por estes dois conceitos). Escolhemos então procurar agentes de referência que nos permitam transformar a população.

Localmente (e relembrando a métrica proposta na eq. 2.1), podemos escrever:

$$\beta = \frac{dx}{x} = d\log x = \log x - \log x_{\mathrm{ref}}, \tag{4.5}$$

onde x_{ref} é o nosso agente de referência: este pode ser um qualquer agente existente ou a média da amostra ou qualquer outro valor dentro da mesma. O algoritmo que iremos utilizar é composto pelos seguintes passos:

1. **Calcular o logaritmo de cada agente:**
   O conceito de processo multiplicativo e a a sua ligação com a distribuição em forma da lei de potência já foram discutidos. Então, faz todo o sentido olhar para a distribuição do logaritmo do agente. Esta transformação tem o beneficio extra de reduzir a variância da distribuição;

2. **Calcular a extensão da população:**
   Fazendo a diferença entre o agente máximo e o mínimo da população obtemos um fator de normalização a que designamos por fator a;

3. **Calcular a distância de cada agente a um agente de referência:**
   A escolha do agente em si não é relevante. Este passo permite uma abstração dos valores absolutos da população, permitindo assim olhar para a distâncias entre agentes. Note-se que este passo permite evitar o problema do que referêncial fixo que é a moeda cunhada;

4. **Normalizar a distância entre agentes:**
   Dividindo a distância de cada agente ao de referência pelo fator a;

5. **Calcular a distância entre dois pontos do espaço-tempo:**
   Este último passo deve-se à expansão do espaço económico e uma justificação, baseada no conceito de redes complexas, é dada no capítulo **A Economia como Rede Complexa**. Note-se que aqui iremos usar um outro agente de referência para subtrair a distância entre agentes causada pela expansão.

[1] https://www.forbes.com/global2000 acedido a 23/maio/2020

No capítulo seguinte iremos aplicar esta transformação à distribuição de uma variável económica e esperamos obter uma distribuição com variância finita.

Capítulo 5

Resultados

5.1 A população

Os dados usados neste estudo são registo de ativos e passivos de empresas portuguesas do ano 2000 a 2010. Por motivos de sigilo bancário, mais nenhuma informação (área de negócios da empresa, anos de atividade, etc.) nos é disponibilizada. Mais especificamente:

- No total, temos cerca de 750 000 registos indexados por número de identificação e ano. Para cada registo temos um valor de ativo e passivo;
- Cerca de 9500 agentes têm registos para todos os anos;
- Há menos de 10 registos que apresentam um dos valores negativos, possivelmente causados por problemas de introdução. Estes registos foram ignorados;
- Para cada par (nº de identificação da empresa, ano em que se verificou essa posição) não há registos duplicados;
- Aproximadamente 1500 registos apresentam valores de passivos ou ativos iguais, registos esses que também foram removidos.

5.2 Normalização da população para um instante

Para começar, olhamos apenas para um ano. A título de exemplo, na Figura 5.1, apresenta-se a distribuição dos ativos em 2007. Observamos a esperada forma de uma lei de potência. Um detalhe interessante que podemos constatar é o facto de a partir de um certo valor de x o gráfico tem o comportamento log-log tal como observado nos dados da WID. Contrariamente aos dados da WID, o lado esquerdo do gráfico

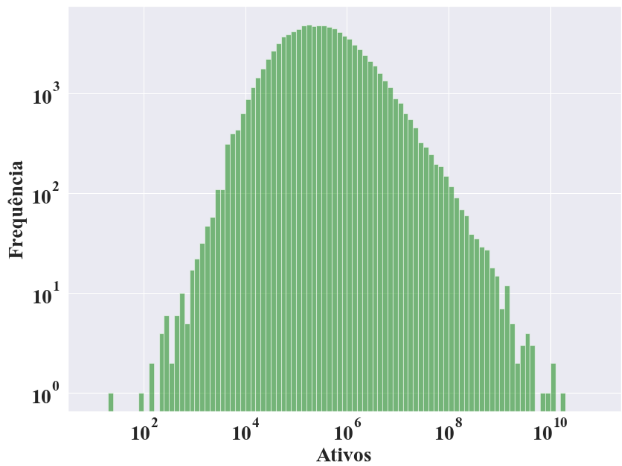

Figura 5.1: Distribuição de ativos de um conjunto de 97877 empresas portuguesas em 2007, gráfico log-log. Os momentos relevantes são apresentados na Tabela 5.1.

decai em vez de se manter horizontal, visto que não existe um conceito equivalente a "salário mínimo nacional" para as empresas.

A Tabela 5.1 apresenta os momentos relevantes desta distribuição. Tal como esperado a variância e excesso de curtose (cujo o valor positivo indica que a distribuição é *heavy-tail*) são várias ordens de grandeza acima do expectável o que dificulta a modelação da população. De facto, os valores desta tabela são obtidos pelas fórmulas usuais mas, tal como já discutido, a forma geral da distribuição significa que os momentos estatísticos não estão definidos, sendo estes valores numéricos uma representação computacional de "infinito".

Tabela 5.1: Momentos da distribuição de ativos em 2007.

Momento	Valor
média	$4,55 \times 10^6$
variância	9×10^{15}
assimetria	104
ex. curtose	14001

Aplicando o algoritmo descrito anteriormente, obtemos a distribuição apresentada na Figura 5.2 e com os momentos da Tabela 5.2.

Um outro resultado interessante é obtido se ordenarmos as distâncias da distribuição (Figura 5.3). A porção central, com menor declive, indica que a maior parte dos agentes estão a uma distancia bem definida do agente de referência, formando uma órbita definida por um poço de potencial.

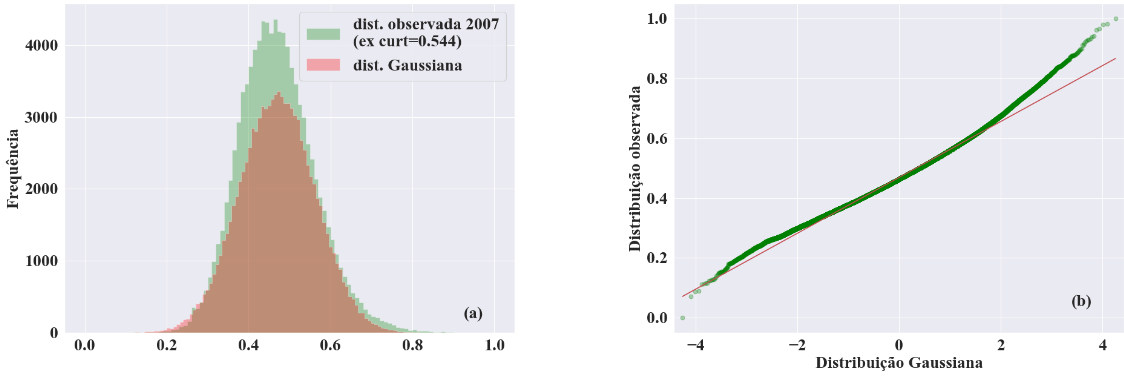

Figura 5.2: Distribuição da distância comóvel para os ativos 2007 (a) juntamente com o gráfico quantil-quantil (b). Sobreposta está uma distribuição Gaussiana com a mesma média e desvio padrão. É também apresentado o gráfico quantil-quantil. Os momentos estatísticos relevantes são apresentados na Tabela 5.2.

Tabela 5.2: Momentos da distribuição de distâncias comóveis de ativos em 2007.

Momento	Valor
média	0,47
variância	0,0087
assimetria	0,41
ex. curtose	0,54

Todos estes resultados são independentes da escolha do agente de referência porque, tal como já foi referido, o sistema económico cresce mas a distribuição de agentes, em média, preserva-se.

Calculando a distribuição para cada ano, conseguimos uma ideia de como evolui o factor de escala (Figura 5.4).

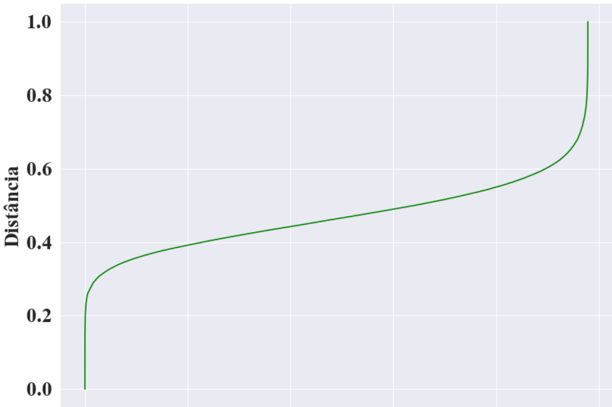

Figura 5.3: Distância comóvel de cada agente ao agente de referência, por ordem crescente. Omite-se o eixo horizontal que contém a identificação numérica de cada agente individual. Observa-se uma zona central com menor declive que indica que a maior parte dos agentes estão a uma distancia bem definida do agente de referência, formando uma órbita definida por um poço de potencial.

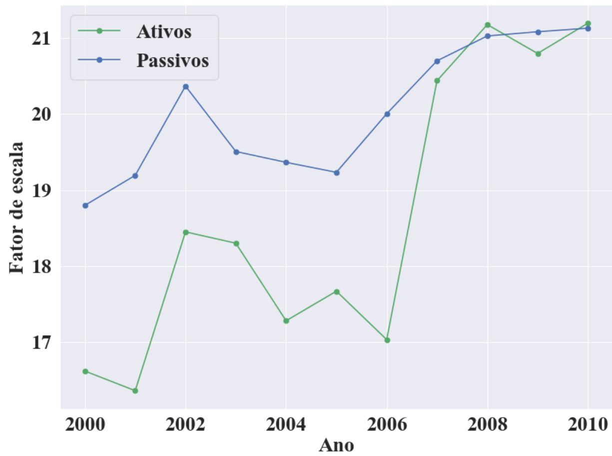

Figura 5.4: Evolução do fator de escala (designado por a) entre 2000 e 2010.

5.3 O fator c

Nesta população, último passo do algoritmo refere-se à normalização entre anos. Para a calcular este fator usamos a equação

$$c = \frac{x_i(\text{ano actual}) - x_i(\text{ano anterior})}{x_i(\text{ano anterior})}, \tag{5.1}$$

onde x_i é o valor de ativos do agente mínimo escolhido. Os resultados estão apresentados na Figura 5.6.

Empiricamente, esperaríamos que o fator c se mantivesse positivo (ver secção 4.1). No entanto, a evolução calculada não reflete isto. Apesar deste comportamento não ser completamente irrealista (basta olhar para a evolução do preço de ações de uma qualquer empresa para perceber que o ganho vem a longo prazo) uma possibilidade é a falta de dados: nada nos garante que os nossos registos incluem sempre o agente mais pobre, ou sequer que o nosso mínimo está correto. Esta dependência de um único registo pode ser mitigada usando o valor da média dos agentes na modelação do fator c. A evolução desta grandeza também inclui informação sobre a evolução da zona morta (ver Figura 4.1). Estudamos então a média dos valores mais baixos (para 10 e 100 agentes) obtemos os resultados da Figura 5.5 e usamos estes três agentes para calcular o fator c (Figura 5.6).

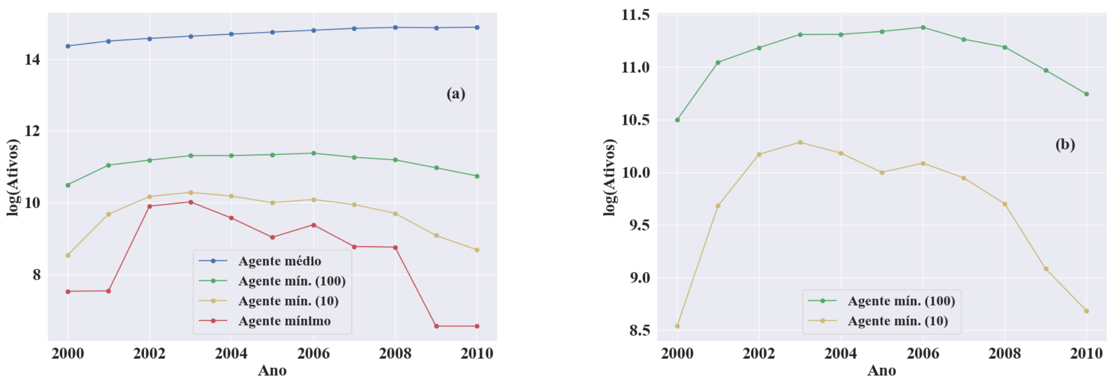

Figura 5.5: Evolução do agente mínimo, em termos de ativos, usando a média global, a média dos 100 agentes mais pobres e 10 mais pobres, assim como do agente mínimo (a) e detalhe para a média dos 100 agentes mais pobres e 10 mais pobres. Em todas as figuras a linha azul representa a média de todos os agentes, a linha verde a média dos 100 agentes mais pobres, a linha amarela os 10 mais pobres e a vermelha o mais pobre (repare-se que de ano para ano os agentes em causa não são necessariamente os mesmos).

Seria também de esperar que a linha cumulativa (Figura 5.6c) fosse monotónica (até o agente mais pobre deveria crescer) mas não é isso que acontece. Uma possível interpretação é que a "linha de água" desceu desde 2003: é necessário menos "esforço" por parte dos agentes para que estes sobrevivam.

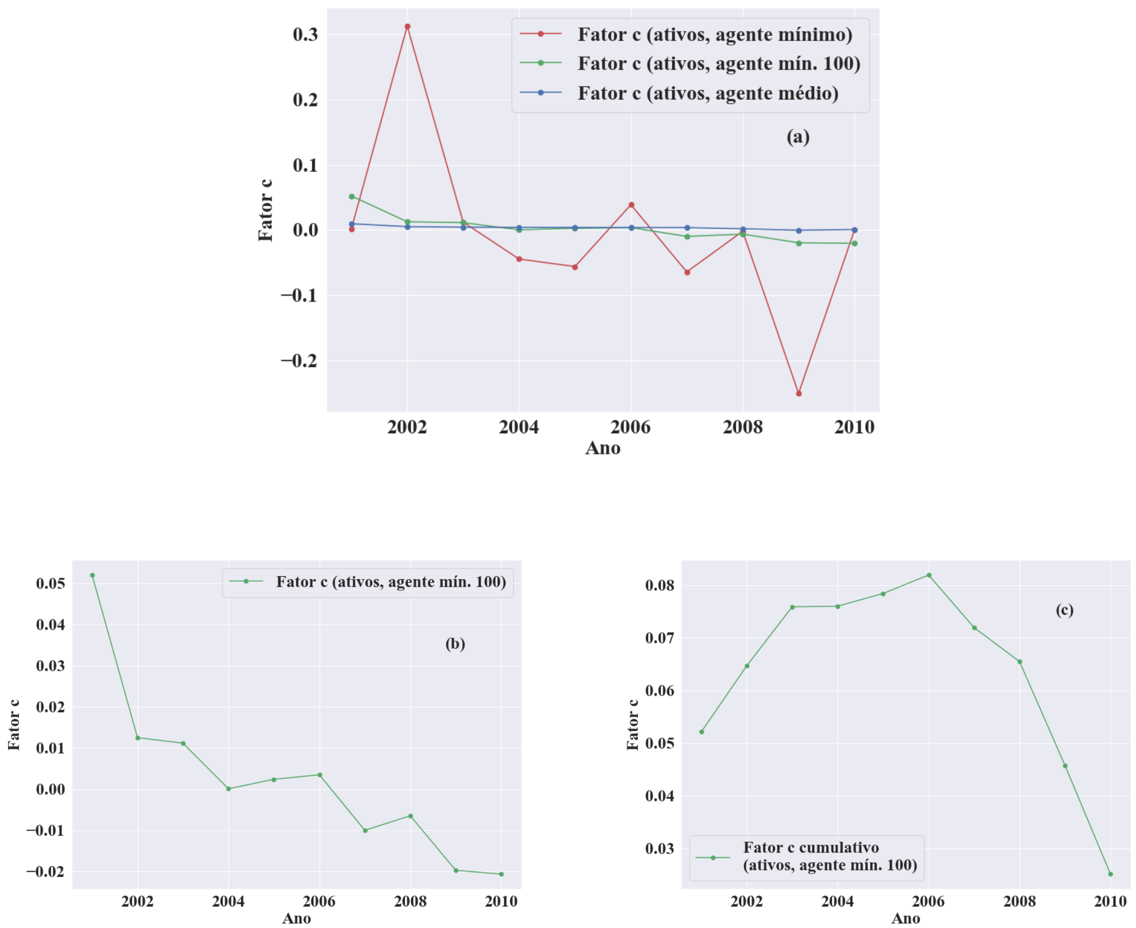

Figura 5.6: Evolução do fator c, em termos de ativos, usando a média global, a média dos 100 agentes mais pobres e o do agente mínimo (a) e detalhe usando a média dos 100 mais pobres (b). Em todas as figuras a linha azul representa o cálculo usando a média de todos os agentes, a linha verde a média dos 100 agentes mais pobres, e a vermelha o mais pobre (repare-se que de ano para ano os agentes em causa não são necessariamente os mesmos).

5.4 População completa

Aplicando o algoritmo de transformação aos dados de cada ano, e usando o fator c como corretor dos dados de diferente anos podemos finalmente juntar todos os registos numa mesma população, obtendo-se a distribuição cujo gráfico está ilustrado na Figura 5.7.

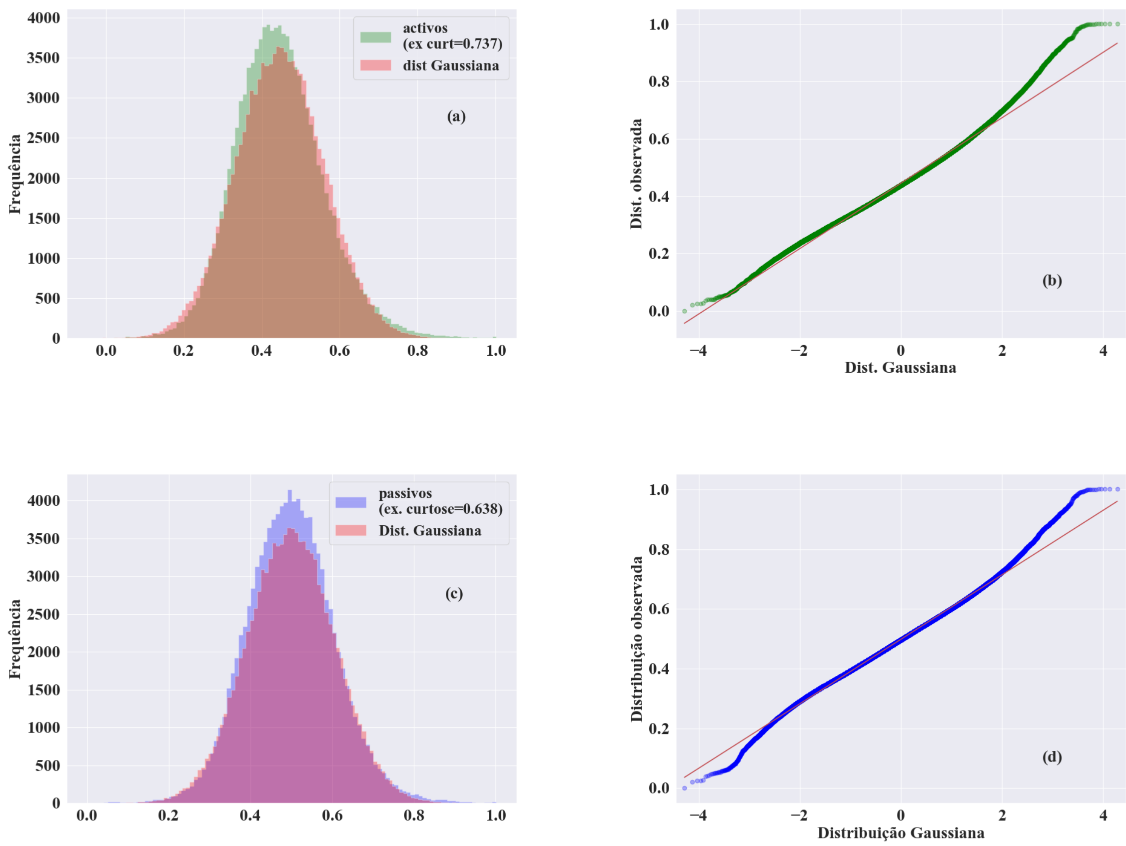

Figura 5.7: Distribuição completa de ativos (a) e passivos (c) usando a média dos 100 agentes mais pobres para calcular o fator c, e os respetivos gráficos quantil-quantil. Nas figuras da esquerda a zona vermelha representa um distribuição Gaussiana com a mesma média e o mesmo desvio padrão que o conjunto das distâncias comóveis.

Capítulo 6

Conclusão

O presente documento estabelece um algoritmo de normalização da distribuição de uma variável económica. De uma forma simples podemos escrever que os dois principais desafios observados são a não Gaussianidade que estas grandezas apresentam e a constante expansão do espaço onde se representam. Para ultrapassar ambos foi necessário beber de duas áreas de conhecimento: o estudo estatístico de varáveis económicas e a teoria de grafos. O primeiro permite uma normalização após a qual o Teorema do Limite Central é aplicável; o segundo é necessário para integrar a expansão do espaço económico na normalização.

Note-se que este algoritmo tem duas limitações cujas consequências podem ser exploradas com mais estatística:

Uma amostra normalizada continua a não ser Gaussiana. Apesar da variância passar a ser finita, ainda não é claro quão válido é assumir que a distribuição terá uma média e desvio padrão bem definidos. Isto é especialmente notório considerando que todas as figuras têm um excesso de curtose positivo. Uma consequência imediata é que valores extremos continuam a ser possíveis (ou , melhor dito, mais prováveis que numa distribuição Gaussiana usual).

Por outro lado, o estudo da expansão do espaço é bastante dependente da qualidade da informação registada para os agentes no limite inferior da variável em estudo. Isto limita a fiabilidade dos resultados que usem medições tomadas em diferentes instantes, especialmente se a fonte de informação ou o método de registo for diferente. Isto, por si só, torna o desenvolvimento desta linha de investigação num projeto a longo prazo: registar a mesma informação, como métodos equivalentes, ao longo de anos.

Ao longo deste trabalho foi tido um especial cuidado na ligação entre a Física e a Economia. Ao progredir numa direção específica (normalização de distribuições com variância infinita) procurou-se estabelecer uma ligação sólida entre as diversas áreas de conhecimento usadas juntamente com bibliografia que ajude a dar os primeiros passos neste cruzamento entre física, economia, estatística.

Um campo onde a estratégia proposta pode encontrar aplicações é o de *machine learning*. Um conceito fundamental (que por vezes é desprezado) nesta área do conhecimento é o quão aprendível é uma população. Um ditado comum é "*garbage in, garbage out*" (quando entra lixo, lixo sai) que se refere ao facto de a aplicação de algoritmos a dados desadequados levará a conclusões erradas. Assim, este algoritmo pode integrar o pré-tratamento de dados económicos para, por exemplo, se a cada registo tiver associada uma etiqueta de vivo/falido poderá ajudar a perceber qual o rácio ideal de ativo/passivo para que um empresa não vá à falência.

Bibliografia

[1] D. J. de Solla Price, “Networks of Scientific Papers”, *Science*, vol. 149, n.º 3683, pp. 510–515, 1965.

[2] A.-L. Barabasi e R. Albert, “Emergence of scaling in random networks”, *Science*, vol. 286, pp. 509–512, 1999.

[3] P. Mirowski, *More Heat than Light: Economics as Social Physics, Physics as Nature’s Economics*. Cambridge University Press, 1989.

[4] E. Smith e D. K. Foley, “Classical thermodynamics and economic general equilibrium theory”, *Journal of Economic Dynamics and Control*, vol. 32, pp. 7–65, 2008.

[5] F. Black e M. Scholes, “The Pricing of Options and Corporate Liabilities”, *Journal of Political Economy*, vol. 81, pp. 637–654, 1973.

[6] A. C. Silva e V. M. Yakovenko, “Temporal evolution of the “thermal” and “superthermal” income classes in the USA during 1983–2001”, *Europhysics Letters (EPL)*, vol. 69, pp. 304–310, 2005.

[7] M. Jagielski e R. Kutner, “Modelling of income distribution in the European Union with the Fokker–Planck equation”, *Physica A: Statistical Mechanics and its Applications*, vol. 392, pp. 2130–2138, 2013.

[8] J. P. da Cruz, H. Cruz, K. Rajaratnam, P. Beling e G. A. Overstreet Jr., “Sand Pile Modeling of Economic Variables For Credit Risk Applications”, 2016.

[9] A. Hoag e J. Hoag, *Introductory Economics*. World Scientific Publishing Co. Pte. Ltd., 2006.

[10] R. Mantegna e H. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, 2000.

[11] N. Derzsy, Z. Néda e M. Santos, “Income distribution patterns from a complete social security database”, *Physica A: Statistical Mechanics and its Applications*, vol. 391, pp. 5611–5619, 2012.

[12] I. Wright, “Implicit Microfoundations for Macroeconomics”, *Economics: The Open-Access, Open-Assessment E-Journal*, vol. 3, p. 1, 2009.

[13] V. Yakovenko e B. Rosser, “Colloquium: Statistical mechanics of money, wealth, and income”, *Reviews of Modern Physics*, vol. 81, pp. 1703–1725, 2009.

[14] V. Pareto, *Manuale d'economia politica*. Droz, Genevae, 1906.

[15] A. B. Downey, "Lognormal and Pareto distributions in the Internet", *Computer Communications*, vol. 28, pp. 790–801, 2005.

[16] K. Park e W. Willinger, *The Internet as a Large-scale Complex System*. Oxford University Press, 2005.

[17] M. Bee, M. Riccaboni e S. Schiavo, "Distribution of City Size: Gibrat, Pareto, Zipf", *The Mathematics of Urban Morphology*, pp. 77–91, 2019.

[18] K. T. Rosen e M. Resnick, "The size distribution of cities: An examination of the Pareto law and primacy", *Journal of Urban Economics*, vol. 8, pp. 165–186, 1980.

[19] V. F. Pisarenko e D. Sornette, "Characterization of the Frequency of Extreme Earthquake Events by the Generalized Pareto Distribution", *Pure and Applied Geophysics*, vol. 160, pp. 2343–2364, 2003.

[20] Y. Y. Kagan, "Earthquake size distribution and earthquake insurance", *Communications in Statistics. Stochastic Models*, vol. 13, pp. 775–797, 1997.

[21] C. Kang, K.-Y. Park e Y.-S. Cho, "Numerical and Statistical Analyses of Tsunami Heights with the L-Moments Method", *Applied Sciences*, vol. 9, p. 5517, 2019.

[22] WID.world. (2019). "World Inequality Database About us", URL: `https://wid.world/wid-world/` (acedido em 23/08/2019).

[23] ——, (2019). "World Inequality Database Generalized Pareto Interpolator", URL: `https://wid.world/gpinter/` (acedido em 23/08/2019).

[24] L. Chancel e A. Gethin, "Building a global income distribution brick by brick", *WID.world Technical Note*, 2017.

[25] F. Alvaredo, A. B. Atkinson, L. Chancel, T. Piketty, E. Saez e G. Zucman, "Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world", *WID.world Technical Note*, jun. de 2017.

[26] D. G. Champernowne, "A Model of Income Distribution", *The Economic Journal*, vol. 63, p. 318, 1953.

[27] T. Blanchett, J. Fournier e T. Piketty, "Generalized Pareto Curves: Theory and Applications", *WID.world Technical Note*, 2017.

[28] L. Euler, "Solutio problematis ad geometriam situs pertinentis", *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, vol. 8, pp. 128–140, 1736.

[29] R. Albert e A.-L. Barabasi, "Statistical mechanics of complex networks", *Reviews of Modern Physics*, vol. 74, n.º 1, pp. 47–97, 2002.

[30] S. N. Dorogovtsev e J. F. F. Mendes, “Evolution of networks”, *Advances in Physics*, vol. 51, pp. 1079–1187, 2002.

[31] M. E. J. Newman, “The structure and function of complex networks”, *SIAM Review*, vol. 45, n.º 2, pp. 167–256, 2003.

[32] D. Sornette, “Multiplicative processes and power laws”, *Physical Review E*, vol. 57, pp. 4811–4813, 1998.

[33] D. Sornette e R. Cont, “Convergent Multiplicative Processes Repelled from Zero: Power Laws and Truncated Power Laws”, *Journal de Physique I*, vol. Vol. 7 / 3 - March 1997, pp. 431–444, 1997.

[34] V. Pareto, *Cours d’Economie Politique. Tome Premier*. V. Giard & E. Brière, 1896.

[35] C. S. Gillespie, “A complete data frame work for fitting power law distributions”, 2014. arXiv: 1408.1554. (acedido em 05/09/2019).

[36] M. Mitzenmacher, “A Brief History of Generative Models for Power Law and Lognormal Distributions”, *Internet Mathematics*, vol. 1, pp. 226–251, 2004.